Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

Seyed Ziaeddin Alborzi [1,3]*, Sabeur Aridhi [1], Marie-Dominique Devignes [2], Rabie Saidi [4], Alexandre Renaux [4], Maria J. Martin [4], David W. Ritchie [3]
1 Universite de Lorraine, LORIA, UMR 7503, 54506 Vandœuvre-les-Nancy, France
2 CNRS, LORIA, UMR 7503, 54506 Vandœuvre-les-Nancy, France.
3 Inria Nancy Grand-Est, 54600 Villers-les-Nancy, France
4 European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK
*To whom correspondence should be addressed: seyed-ziaeddin.alborzi@inria.fr

1. INTRODUCTION

The GO ontology is widely used for functional annotation of genes and proteins. It describes biological processes (BP), molecular function (MF), and cellular components (CC) in three distinct hierarchical controlled vocabularies. At the molecular level, functions are often performed by highly conserved parts of proteins, identified by sequence or structure alignments and classified into domains or families (SCOP, CATH, PFAM, TIGRFAMs, etc.). The InterPro database provides a valuable integrated classification of protein sequences and domains which is linked to nearly all existing other classifications. Interestingly, several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only 20% of Pfam domains and families possess MF GO functional annotation). We therefore developed the GODM approach to expand the available functional annotations of protein domains and families (1). Based on our ECDomainMiner approach (2), we use the respective associations of protein sequences with GO terms and protein domains to infer direct associations between GO terms and protein domains.

2. INFERRING GO-DOMAIN ASSOCIATIONS USING GODM

GODM finds associations between GO terms and protein domains from the known associations between (i) GO terms and protein sequences and (ii) the same protein sequences and the domains they are known to contain. The domains may belong to any domain classification such as Pfam. We used two types of datasets: (i) SIFTS for associations between PDB chains and GO terms or domains, (ii) the Swissprot and TrEMBL sections of UniProtKB for associations between sequence accession numbers (ANs) and GO terms or domains. Next, based on the evidence code of the GO term assignment, AN-GO term associations in the SwissProt and TrEMBL datasets are divided into two groups, namely associations for which GO terms were Inferred from Electronic Annotation (IEA) and the rest. These four input datasets are subsequently called Swissprot, Swissprot-IEA, TrEMBL, and TrEMBL-IEA. In order to exploit the GO hierarchy, associations involving ancestors of GO terms are also added to the datasets. Finally, PDB chains and ANs are grouped into non-redundant clusters having identical sequences using the Uniref100 resource.

In each dataset prepared in this way, each GO term and domain is assigned a feature vector of associated chain or AN clusters. This allows to calculate cosine similarities between GO terms and domains. The scores assigned to each vector pair in each of the five datasets are combined using a weighted average. The individual weights are optimised by calculating the ROC performance plot and maximizing the AUC with manually confirmed GO-Domain associations from InterPro as positive examples, against all others. Then, a threshold is chosen for the weighted score in order to eliminate weak GO-domain associations. Finally a *p-value* is calculated for each GO-domain association in each dataset using a hypergeometric distribution.

3. RESULTS FOR GO-PFAM ASSOCIATIONS

The GODM method infers 20,318 GO-Pfam associations where GO terms are leaves in the MF hierarchy of GO terms. Compared to the 1561 manually curated GO-Pfam associations in InterPro, this represents a 13-fold increase in the number of GO-Pfam associations. Furthermore, the GODM associations have been compared with the dcGO database (3) that includes 3,086 comparable GO-Pfam associations. A total of 2,401 GO-Pfam associations are common between dcGO and our results revealing that our GODM dataset contains 17,917 additional GO-Pfam associations. Moreover the overlap with the 1561 InterPro GO-Pfam associations is of 1519 for the GODM dataset versus only 404 for the dcGO dataset. The GODM method was also run with the SCOP and CATH classifications of domains or families and yielded very similar results.

4. USING THE GODM RESOURCE TO GENERATE ANNOTATION RULES

In this section, we present a systematic way to generate high confidence rules for protein annotation using the GODM associations. We first ran GODM several times to find associations between GO terms and domains from the different domain classifications (such as PFAM, TIGRFAMs, etc.). Then, all associations were grouped for each given GO term resulting in an association of the GO term with a set of domains pertaining from diverse classifications. We then generated all possible subsets of domains ($\{D_1,..., D_n\}$, $n \leq 4$) and associated them with the concerned GO term, GO_k. The subsets of domains were further diversified by adding a taxon (T_j) from a list of interest (one per subset). These complex associations, ($\{\{D_1,..., D_n\}, T_j\}, Go_k$), were converted into annotation rules:

IF a sequence S belongs to taxon T_j and S contains domains $\{D_1,..., D_n\}$
THEN S is annotated by GO_k.

In order to verify the quality of each generated rule, a confidence score was assigned as the ratio of the number of SwissProt sequences verifying the rule over the number of SwissProt sequences verifying the premise of the rule. Candidate rules with high confidence (usually 100%) are retained and used to assign GO terms to unannotated protein sequences. When using Pfam, SCOP, CATH, Panther, PROSITE, CDD, SMART, PRINTS, and TIGRFAM domain classification for GODM, and a set of 40 taxa from CAFA3 unannotated protein sequences, we obtained 6,357, 17,466, and 2,338 annotation rules for MF, BP, and CC GO terms with 100% confidence on SwissProt. These rules were used to annotate target protein sequences in the CAFA3 challenge (http://biofunctionprediction.org/cafa/). There were a total of 121,914 target sequences having at least one known domain present in our GODM-derived rules. Using our high confidence annotation rules, we obtained 188,549 MF, 315,310 BP, and 191,835 CC GO term predictions for 98,849, 106,346, and 105,274 distinct CAFA3 target sequences, respectively.

5. CONCLUSION

The GODM approach provides a substantial enrichment of functional annotations at the protein domain level which has been exploited here for protein functional annotation but can also be used to deepen our knowledge about structure-function relationships at the domain level.

6. REFERENCES

1. Alborzi, S.Z., Devignes, M.D. and Ritchie, D.W., 2017, April. Associating Gene Ontology Terms with Pfam Protein Domains. *In International Conference on Bioinformatics and Biomedical Engineering* (pp. 127-138). Springer, Cham.
2. Alborzi, S.Z., Devignes, M.D. and Ritchie, D.W., 2017. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC bioinformatics*, 18(1), p.107.
3. Fang, H. and Gough, J., 2013. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic acids research*, 41(D1), D536-D544.